Hybrid Text Mining Approach to Prediction of Customer Behaviour in Online Shopping

gkbZfCkzM VSDLV ekbZfuax vijksp Vw ijfMd'ku vkWQ dLVej fcgsfo;j bu vksuykbZZu 'kkWfiax

A Synopsis

Submitted in partial fulfilment of registration in Doctor of Philosophy In Computer Science and Applications

Submitted by:

Ram Gopal

Roll No: 15149740005

Under the Supervision of

Dr. Vikram Singh

Department of Computer Science & Applications

Chaudhary Devi Lal University, Sirsa (Haryana)

Content

Page No.

1. **INTRODUCTION** **1**

1.1. DATA MINING **1**

1.1.1. TEXT MINING 2

1.2. WEB LOG MINING **5**

1.2.1. Web Content Mining (WCM) 5

1.2.2. Web Structure Mining (WSM) 5

1.2.3. Web Usage Mining (WUM) 6

2. **REVIEW OF LITERATURE** **6**

3. **RESEARCH OBJECTIVES AND METHODOLOGY** **11**

3.1. PROBLEM STATEMENT **11**

3.2. RESEARCH OBJECTIVES **12**

3.3. RESEARCH METHODOLOGY **12**

REFERENCES

INTRODUCTION

Modern era is the era of computers and now a days due to emergence of media everything is becoming online. People are using smart phone so many social web sites are being used for interaction with each other and also many web sites are used for buying goods online now a day we can buy everything online and people are using internet 24x7 hours. Hence everyone is in the addiction of web now a days.

DATA MINING

Data Mining is an interdisciplinary field of Computer Science. It is an umbrella (Artificial Intelligence, statistic, mathematics, neural networks, database etc.) of techniques that discovers previously unknown relationship in data. The overall goal of the data mining process is to extract information from a data set and transform it into an understandable structure for future use, **Ahmed R.A. et. al, (2015)** Data mining is crucial for extracting and identifying useful information from a large amount of data.

Data mining techniques can be classified broadly as, **Harshni S. et. all, (2017),**

1. Predictive Classification, Regression, Time series Analysis, Prediction.
2. Descriptive Clustering, Summarization, Association Rules, Sequence Discovery

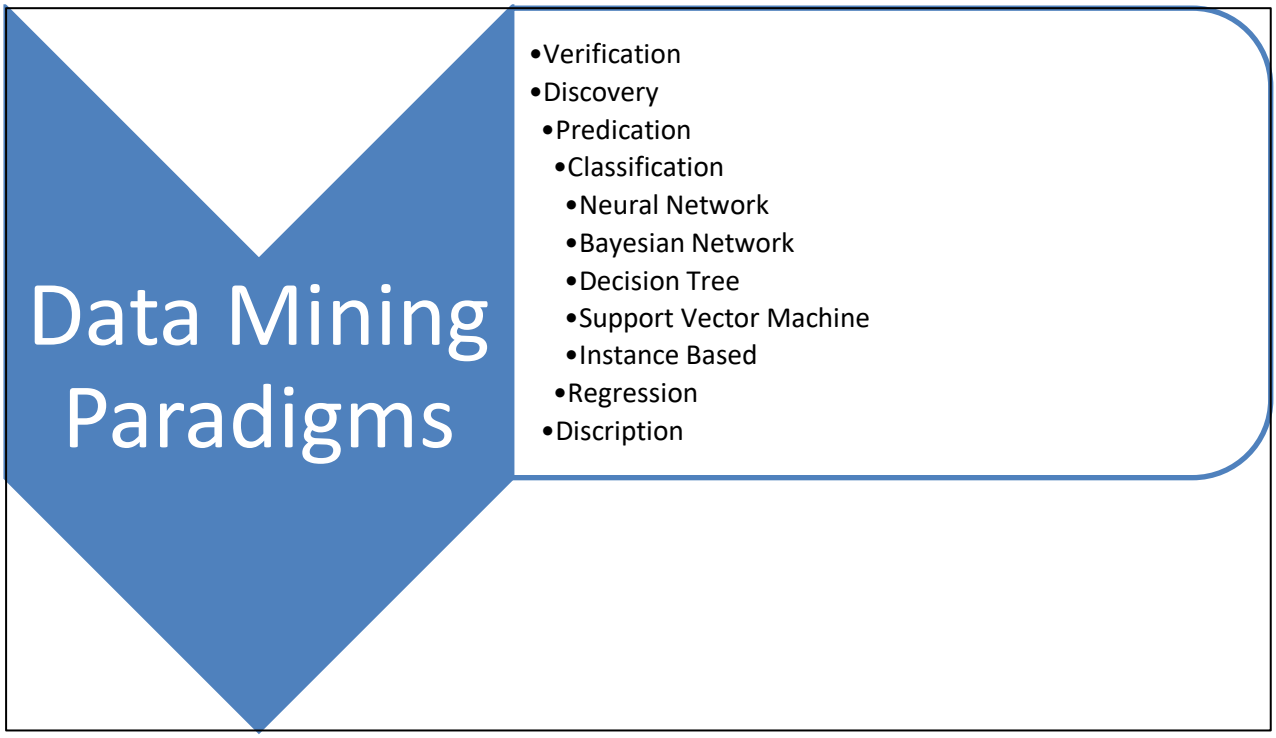

Fig 1. Data Mining Pradigm Classification

Data Mining Process :

1. Selection: This is a process of selecting relevant data from database or data warehouse which are useful for data mining.
2. Pre-processing and cleaning: Raw data we used to collect are not always clean and may contain errors, missing values, noisy or inconsistent data. Thereby getting rid of such anomalies is an important task to achieve.
3. Features selection and extraction: Redefining the data with a smaller number of attributes than the original set.
4. Data Mining: Here various algorithms are applied on the data for getting patterns from the parameters so that for future knowledge can be extracted.
5. Interpretation and Evaluation: This is where we generate visualization, transformation, forecasting and predictions.
   (www.tutorialspoint.com/data_mining/dm_quick_guide.htm)

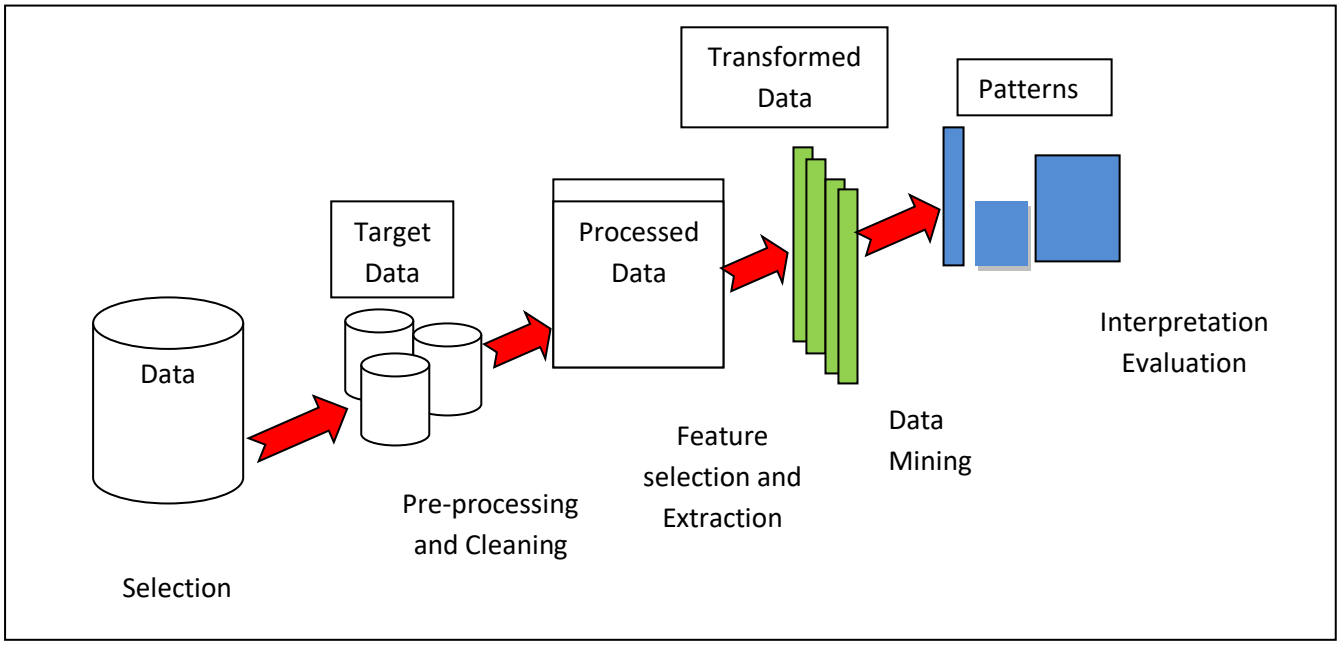

Fig 2 Stages in Data Mining

TEXT MINING

Harshni S. et. all, (2017), The main reason for using data mining is to collect the needed and useful information from vast and enormous amount of data. The sub domain used in extracting the unique text is called as text mining. This is used to retrieve the main texts from unstructured and semi-structured textual formats. Generally all data in the web and social media are available in the form of fuzzy and random manner. So the main aim is to establish a bridge and relationship among the texts, so that it becomes easier for human to understand and process it in a effective manner. This process is

known as Knowledge discovery from texts(KDT). **Ertek G. et. all, (2017),**Text mining is the branch of data mining aimed at discovering interesting and nontrivial knowledge from text documents. The common practice in text mining is the extraction of critical attribute information from unstructured text documents through text processing, and then analyzing this structured constructed data with well-known data mining algorithms. Attribute selection allows the removal of the irrelevant and noisy information present in text documents and focuses to only on relevant and informative data for use in text mining.

Harshni S. et. all, (2017), The process of text mining includes resolving useful patterns from unstructured and semi-structured data. Computers have low capability when compared to humans in analyzing linguistic patterns, but processing can be done at a higher speed. So large amount of algorithms can be used to analyze and process them. The main functions of text mining are: (1)Text summarization (2)Text categorization (3)Text clustering

ALGORITHMS:

Apriori algorithm

Ravichandran S. et. all, (2012), The Apriori Algorithm is an prominent algorithm for mining frequent item sets for association rules. Apriori is designed to operate on databases containing transactions (for example, collections of queries from customers, or details of a website frequentation). Classification using association rules combines association rule mining and classification. It is concerned with finding rules that accurately predict a single solution. In association rule mining, various new rules found can be updated in Rule base. Apriori Algorithm is suitable to compute the rules and patterns and predict for any organization to improve the customer satisfaction. An association rule is "strong" if its confidence value is greater than a user-defind threshold. The association rules are created by combining each large item set with each of its subsets. The Apriori algorithm is an efficient algorithm for finding all frequent item sets. It implements level-wise search using frequent item property and can be additionally optimized.

FP tree algorithm

Suguna S. et all, (2016), FP growth algorithm (Frequent pattern growth algorithm) is an improvement of apriori algorithm. FP growth algorithm used for finding frequent item-set in a transaction database without candidate

generation. It is Faster than apriori algorithm, No candidate generation, Only two passes over dataset.

Agglomerative Algorithm

Agglomerative hierarchical clustering is a bottom-up clustering method where clusters have sub-clusters, which in turn have sub-clusters, etc. The classic example of this is species taxonomy. Gene expression data might also exhibit this hierarchical quality (e.g. neurotransmitter gene families). Agglomerative hierarchical clustering starts with every single object (gene or sample) in a single cluster. Then, in each successive iteration, it agglomerates (merges) the closest pair of clusters by satisfying some similarity criteria, until all of the data is in one cluster.

The hierarchy within the final cluster has the following properties:

- Clusters generated in early stages are nested in those generated in later stages.
- Clusters with different sizes in the tree can be valuable for discovery.

Naive Bayes Classifier

Cheriyan S.,Chitra K.,(2013), It is a classification technique based on Bayes' Theorem with an assumption of independence among predictors. In simple terms, a Naive Bayes classifier assumes that the presence of a particular feature in a class is unrelated to the presence of any other feature. For example, a fruit may be considered to be an apple if it is red, round, and about 3 inches in diameter. Even if these features depend on each other or upon the existence of the other features, all of these properties independently contribute to the probability that this fruit is an apple and that is why it is known as 'Naive'.

K-means Clustering algorithm

Yadav M., (2012), K-means is simplest unsupervised learning algorithms that solve the clustering problem. It classify a given data set through a certain number of clusters (assume k clusters) fixed apriori. The main idea is to define k centers, one for each cluster. These centers should be placed in a cunning way because of different location causes different result. So, the better choice is to place them as much as possible far away from each other. The next step is to take each point belonging to a given data set and associate it to the nearest center. When no point is pending, the first step is

completed and an early group age is done. At this point we need to re-calculate k new centroids as barycenter of the clusters resulting from the previous step. After we have these k new centroids, a new binding has to be done between the same data set points and the nearest new center. A loop has been generated. As a result of this loop we may notice that the k centers change their location step by step until no more changes are done or in other words centers do not move any more.

WEB LOG MINING

In modern Era there is an emergence of media and people are connected with each other on the World Wide Web via various hand held or computing devices. E-marketing is also showing its adverb effect in our day to day life web consists of a massive collection of textual documents called web pages. Customers are purchasing things online, hence it's the basic need of e-marketing to predict the behaviour of the customer to better understand the customer requirements for awesome business and customer relations. With the explosive growth of information sources available on the WWW, it has become an important tool for users in order to find, extract, filter and evaluate the desired information and resources, so various methods are used for mining these user interactions with the web servers. For this raw data which is in the form of logs on the servers is taken and valuable information from that data is retrieved and the process of mining the web data is known as web mining. The process of web mining is divided into four stages source data collection, data pre-processing, and pattern discovery and pattern analysis.

Cheriyan S.,Chitra K.,(2013), Three types of web mining are

Web Content Mining (WCM)

Uncovering of useful information from web contents including text, image, audio, video etc. research in web content mining encompasses resource discovery from web, document categorization and clustering and information extraction from web pages.

Web Structure Mining (WSM)

Web structure mining studies the web's hyperlink structure. It usually involves analysis of the in-links and out-links of a web page, and it has been used for search engine result ranking.

Web Usage Mining (WUM)

It focuses on analyzing search logs or other activity logs to find interesting patterns. One of the main application of web usage mining is to learn user profiles.

ROLE OF TEXT MINING IN ONLINE SHOPING

Improving the design of e-commerce site, Security system for intrusion detection or automatic request generating programmes for saving a web server from bottle neck problem, Understanding customer needs for future plans like advertisement of product demands of products broadly predication of customer behaviour leads to the exponential growth in the e-marketing hence with the help of text mining techniques customer needs can be figured out and can be full filled in the boundaries of time for batter customer relationships.

REVIEW OF LITERATURE

Hernadez S. et. al., (2016) have proposes a linear-temporal logic model checking approach for the analysis of structured e-commerce web logs. By defining a common way of mapping log records according to the e-commerce structure, web logs can be easily converted into event logs where the behaviour of users is captured. Then, different predefined queries can be performed to identify different behavioural patterns that consider the different actions performed by a user during a session. Finally, the usefulness of the proposed approach has been studied by applying it to a real case study of a Spanish e-commerce website.

Luo X. et. al., (2017), have Introduced the method of web log mining, which can discover the mode of web pages by digging web log records. By analyzing and exploring the rules of web log records, potential customers of the website can be identified and quality of information services to users can be improved. In the stage of user behaviour analysis, they explored the differences in user browsing behaviour in different types of access events, and calculated the user's interest based on the M5 model tree to analyze the analytic events.

Volk M et. al.,(2017), A comprehensive workflow is proposed using a clustering algorithm in a Hadoop ecosystem to investigate user interest patterns. The complete workflow is demonstrated on an application scenario of one of the largest business-to-business (B2B) electronic commerce websites in Germany. Furthermore, an experimental evaluation method is applied to verify the applicability and efficiency of the used algorithm, along with the associated framework.

Hsueh N.L., (2016), Has presented a framework for opinion mining in Traditional Chinese- called FOM (Framework of Opinion Mining) to collect unstructured articles in the popular web site and analyse the opinion and sentiment in the semiautomatic way. The framework is developed by objected oriented design patterns, such as to support the flexibility and maintainability. With the FOM framework, new analysis algorithm can be easily replaced and integrated in a new application. A flood predication application based on facebook text in Taiwan is demonstrated.

Bodendorf F. et. all,(2010), have presented opinion mining approach which allows an automated extraction, aggregation, and analysis of customer opinions on products by using text mining. Thus, strengths and weaknesses judged by customers can be detected at an early stage and starting points for product design and marketing can be identified. The application of the approach is illustrated by a case study coming from the automotive industry.

Hemapriya A. et. al., (2015), proposed a modified version of link based heuristic to handle back browser movements. It also made use of the time heuristics. The proposed system is compared with an existing system that applies links heuristics for consecutive records and is found to be better.

Yadav M., (2012),has studied the customer's behaviour using the Web mining techniques and its application in e-commerce to mine customer behaviour. The concept of Web mining describing the process of Web data mining in detail: source data collection, data preprocessing, pattern discovery, pattern analysis and cluster analysis. With the advanced information technologies, server are now able to collect and store mountains of data, describing their numerous contributions and different customer profiles, from which they seek to derive information about their customer's requirements. Conventional methods are no longer appropriate for these business

situations to find the customer behaviour. The principle of data mining is to cluster customer segments by using K-Means algorithm in which input data comes from web log of various e-commerce websites. Hence, determine the relationship between Web data mining and ecommerce and also to apply Web mining technology in ecommerce.

Ahmed R., (2015), eleven data mining classification techniques will be comparatively tested to find the best classifier fit for consumer online shopping attitudes and behavior according to obtained dataset for big agency of online shopping ,the results shows that decision table classifier and filtered classifier gives the highest accuracy and the lowest accuracy is achieved by classification via clustering and simple cart, also this paper will provide a recommender system based on decision table classifier helping the customer to find the products he/she is searching for in some ecommerce web sites .Recommender system learns from the information about customers and products and provides appropriate personalized recommendations to customers to find the desired products.

Chen L, Su Q.,(2013), has proposed new definition of users' interest for the first time as a set of the preference for commodity categories. In order to describe users' behaviours and reflect their interest, three main indicators category visiting path, browsing frequency and relative length of access time are taken into consideration and refined from clickstream data. According to these indicators, an improved clustering algorithm with rough set theory is used to cluster users with similar interest. The experimental result shows that this algorithm is effective and applicable.

Facca F.,(2005), have presented a survey of the recent developments in the area that is receiving increasing attention from the Data Mining community. Reference between applications, techniques, and data sources is also found.

MANFRED M. et. all.,(2011), have suggested AdMotional aiming at achieving a win-win situation for online advertisers and web users alike through optimization of the campaign selection process and the creation of individually personalized ads. This approach is expected to result in increased campaign performance for advertisers, and in more relevant – and thus less annoying – ads for consumers. A general overview of the system architecture before describing the main components in detail. For

campaign selection, AdMotional combines and enhances state-of-the-art targeting technologies by a novel concept of emotional targeting. The key aspect of ad personalization furthermore demands for the dynamic creation of the ad itself for which different approaches are presented and evaluated. We then introduce the learning and optimization component and strategies before concluding with a summary and brief outlook into future developments.

Al-Yazeed M.A., (2015), has proposed a hybrid page ranking model based on web usage mining technique by exploiting session data of users, to enhance the recommendations of the next candidate web page to be accessed. The proposed approach represented a combination between two page ranking approaches. The first one computes the frequency ratio indicating the number of occurrences of each page in the search result. On the other hand, the second approach computes the coverage ratio from similar behaviour patterns. As a result of the proposed approach, a set of candidate pages are ranked and the page with highest rate is recommended. The proposed approach has been tested on real data collected and extracted from the web server log file of CTI main web server.

Lavanya T. Et. all, (2016), have designed an algorithm which extracts opinion targets and opinion words using word alignment model for online reviews extracted from Twitter. An opinion target is defined as the topic about which users express their opinions. An opinion words are defined as the words that are used to express users' opinions. The aim of the project was to determine the thoughts of blog or review writer with respect to some topic or the overall contextual polarity of online reviews using word alignment model. The key objective of this project was to design an algorithm that predict opinion words and opinion targets for analyzing the market status of a product by mining user reviews posted in social networking site namely the Twitter. For evaluation, Benchmark Customer Review Dataset and real tweets dataset pertaining to product reviews extracted using Twitter API. The results are measured in terms of Precision and Recall for accuracy of finding opinion words and targets in order to be an essential ingredient for opinion mining and sentiment analysis.

Cheriyan S. and Chitra K.,(2013), have adopted Markov model which is assumed to be a probability model by which users' browsing behaviours can be predicted at category level. Bayesian statistics can also be applied to present and infer users'

browsing behaviours at webpage level. In this research, Markov models and Bayesian theorem are combined and a two-level prediction model is designed. By the Markov Model, the system can effectively filter the possible category of the websites and Bayesian theorem will help to predict websites accuracy.

Pushpalatha N., Reddy S., (2017), have presented a framework named eXtensible Web Usage Mining Framework (XWUMF) for discovering actionable knowledge from web log data. The framework employs a hybrid approach that exploits fuzzy clustering methods and methods for user behaviour analysis. Moreover the framework is extensible as it can accommodate new algorithms for fuzzy clustering and user behaviour analysis. they proposed an algorithm known as Sequential Web Usage Miner (SWUM) for efficient mining of web usage patterns from different datasets. They built a prototype application to validate the framework.

Ravichandran S. et. all, (2012), Have proposed a model that integrates the customer queries, transactions, databases and all other specifications used in ERP Systems, then use data mining techniques to integrate decision making and forecasting. Using ERP characteristics, data gathered from central database are in cluster format which is based on action taken against the queries generated by customers. Later the clustered data's are used by Apriori algorithm to extract new rules and patterns for the enhancement of an organisation.

Suguna S. et all, (2016), have discussed about the importance of log files in E-commerce world. The analysis of log files is used for learning the user behavior in E-commerce system. The analysis of such large web log files need parallel processing and reliable data storage system. The Hadoop framework provides reliable storage by Hadoop Distributed File System and parallel processing system for large database using MapReduce programming model. These mechanisms help to process log data in parallel manner and computes results efficiently. This approach reduces the response time as well as load on the end system. This work proposes apredictive prefetching system based on preprocessing of web logs using HadoopMapReduce.

Tsumoto S. et all, (2017), Have focused on text mining for discharge summaries, which include the compact explanation for the patient’s admission. a record of her complaints, physical findings, laboratory results and radiographic studies while

hospitalized; a list of changes in her medications at discharge; and recommendations for follow up care. Text mining process consists of the following four processes: first, morphological analysis is applied to a set of summaries and a term matrix is generated. Second, correspond analysis is applied to the classification labels and the term matrix and generates two dimensional coordinates. By measuring the distances between categories and the assigned points, ranking of key words will be generated. Then, keywords are selected as attributes according to the rank, and training examples for classifiers will be generated. Finally, learning methods are applied to the training examples. Experimental validation shows that random forest achieved the best performance and the second best was the deep learner with a small difference , but decision tree methods with many keywords performed only a little worse than neural network or deep learning methods.

Cusumano J.L.G. et. all, (2017), observed that Correspondence analysis (CA), SVD plot and crosscorrelation analysis (CCA) are used to analyze concepts and words, correlated in time and to create a warning/forecast mechanism. They apply CA and CCA to Toyota's quality control problems related sudden acceleration of their cars. The predictive model a "cross correlated autoregressive" created helps to determine the number of lags (weeks) which could have been used for the detection of Toyota Accidents.

Manfred Meyer et al. (2016), Present AdMotional to enable online advertisers and web users for optimization of compaign selection process and creation of individually personalized ads. The key aspect of ad personalization furthermore demands for the dynamic creation of the ad itself for which different approaches are presented and evaluated. It utilizes established targeting strategies for its selection process and also exploits potential in the personalization of template-based ads, enhancing personalization.

RESEARCH OBJECTIVES AND METHODOLOGY

PROBLEM STATEMENT

Due to exponential increase in online shopping , for enhancement of business customer relation management is very essential to design websites according to user

requirement and also track the user activities for predication of future purchases to control the production and demand factor. This predication can be done with the help of server log files which records activity of website usages.With the help of log files inforrmation of Browser and operating system used or PC or mobile phone used can be discovered. Design of e-commerce sites can be improved, Security system for intrusion detection or detection of automatic request generating programs for saving a web server from bottle neck problem or DOS attack. Understanding customer needs for future plans like advertisement of product demands of products etc.

M5 algorithm and many other algorithms are there but many algorithms are slow and some scans data sets many times or some algorithms are memory killer. Some algorithms gives us useless patterns which are not needed so an algorithm should be their which will provide patterns on user requirements. To bridge the gap between current trends in e-markets and our already existing algorithms for better customer relations. user Predication of customer behaviour in e-marketing for future purchase as well as for demand and production estimates in online shopping is need to be study. To increase the performance of web sites better web site design, web server activities are changed as per users' interests. To achieve this We have to analyze user access pattern which are captured in the form of log files.

These factors motivated the research to do research in this direction.

RESEARCH OBJECTIVES

1. To study existing Algorithm for prediction of customer behaviour in online shopping. (K-means, apriori, decision tree, FP Tree)
2. To propose a hybrid algorithm for prediction of customer behaviour in online shopping.
3. To evaluate the performance of the proposed algorithm and compare it with existing algorithms for time, space etc.

RESEARCH METHODOLOGY

Data will be taken from web servers log file and cookies. Then these web log files and cookies will be cleaned and then this unstructured data will be structured for pattern generation and then on the basis of generated patterns decision can be taken

our approach will be tested comparatively with all ready existing algorithms on the basis of time and space.

Some commercial and free open source web mining softwares are

1. WebLog Expert
2. Click Tracks
3. Amadea Web Mining
4. Nihuo Web Log Analyzer
5. Surf Pattern Visual Analyzer
6. ANGOSS KnowledgeWebMiner
7. Htminer etc.

REFERENCES:

Hernadez S,P, Javier Fabra A. and Ezpeleta J. (2016), "Analysis of user's Behaviour in structured e-commerce websites". *2016-IEEE.translation and content mining are permitted for academic research only, Personal use is also permitted, But republication/redistribution require IEEE permission, 2169-3536.*

Luo , X et. al., (2017), "User Behaviour Analysis Based on User Interest by Web Log Mining".,*IEEE 2017 27th International Telecommunication Network and Applications Conference (ITNAC)*, 978-1-5090-6796-1/17 © 2017

Volk M. et. al.,(2017). "New E-Commerce user Interest Patterns" , *Germany 2017 IEEE 6th International Conference on Big Data*.

Hsueh N.L., (2016), "A Framework for Opinion Mining System with Design Pattern", *2016 International Computer Symposium*, 978-1-5090-3438-3/16.

. Bodendorf F. et. all,(2010), "Mining Customer Opinions on the Internet A Case Study in the Automotive Industry", *2010 Third International Conference on Knowledge Discovery and Data Mining*, 978-0-7695-3923-2/10 $26.00 © 2010 IEEE DOI 10.1109/WKDD.2010.129.

Hemapriya A. et. al., (2015),,"Reconstruction of User Sessions from Web Server Log*", ©2015 IEEE,. 978-1-4799-7623-2/15*

Yadav M.P. et. al.,(2012), "Mining the customer behaviour using web usage mining in e-Commerce". *IEEE-20180,ICCNT 12, 26th 28th july 2012, Coimbatore, India*.

Ahmed R.A. at. el., (2015), "Performance study of classification algorithms for consumer online shopping attitudes and behaviour using data mining". *978-1-4799-1797-6/15 © 2015 IEEE, DOI 10.1109/CSNT 2015 50. 2015 Fifth International conference on Communication Systems and Network Technologies.*

Chen L, Su Q.,(2013), " Discovering User's Interest at E-Commerce site Using Clickstream Data". 978-1-4673-4843-0/13© 2013 IEEE.

Facca F.M., and Lanzi P.L,(2005), "Mining interesting knowledge from weblogs: A survey". Data and Knowledge engineering Vol. 53 no. 3, pp 225-241,2005.

MANFRED M. et. all.,(2011), "ADMOTIONAL: TOWARDS PERSONALIZED ONLINE ADS" International Journal of Computer Science and Applications,, Technomathematics Research Foundation, Vol. 8, No. 2, pp. 59 – 80, 2011.

Al-Yazeed M.A., (2015), "A Hybrid Recommendation model for web navigation.", 2015 IEEE seventh international conference on intelligent computing and information systems.

Lavanya T. Et. all, (2016) "Online Review Analytics using Word Alignment Model on Twitter Data", 2016 3rd International Conference on Advanced Computing and Communication Systems (ICACCS -2016), Jan. 22 & 23, 2016, Coimbatore, INDIA.

Cheriyan S.,Chitra K.,(2013)," Web page prediction using Markov Model and Bayesian Statistics"

Pushpalatha N., Reddy S., (2017), "Towards an Extensible Web Usage Mining Framework for Actionable Knowledge", International Conference on Inventive Communication and Computational Technologies (ICICCT 2017).

Ravichandran S. et. all, (2012),"RULE-BASE DATA MINING SYSTEMS FOR CUSTOMER QUERIES", ICCCNT'12 26t _28t July 2012, Coimbatore, India.

Suguna S. et all, (2016), "Big Data Analysis in E-commerce System Using HadoopMapReduce"

Tsumoto S. et all, (2017), “Mining Text for Disease Diagnosis in Hospital Information Syste”, 2017 IEEE International Conference on Big Data (BIGDATA).

Cusumano J.L.G. et. all, (2017), “A Detection Mechanism with Text Mining Cross Correlation Approach”, 2017 IEEE International Conference on Big Data (BIGDATA).

Ertek G. et. all, (2017),” Text Mining Analysis of Wind Turbine Accidents: An Ontology-Based Framework”, 2017 IEEE International Conference on Big Data (BIGDATA).

Yang D., Hong J.,(2017),” Performing literature review using text mining, Part II: Expanding domain knowledge with abbreviation identification”, 2017 IEEE International Conference on Big Data (BIGDATA).

Putra S. et. all,(2016),” Text Mining for Indonesian Translation of the Quran: A Systematic Review”

Miao D.,Fei L., (2017),” A Recommendation System Based on Text Mining”, 2017 International Conference on Cyber-Enabled Distributed Computing and Knowledge Discovery.

Palshikar G. et. all,(2017), “HiSPEED: A System for Mining Performance Appraisal Data and Text”, 2017 International Conference on Data Science and Advanced Analytics., 978-1-5090-5004-8/17 $31.00 © 2017 IEEE DOI 10.1109/DSAA.2017.45.

Wang L. et. all,(2016),” Text Mining–based Evaluation of the User Experience in Online Shopping for Clothing”, This research is supported by funding from China’s central university basic scientific research expenses, Project No. N152304008, and by funding from Northeastern University at Qinhuangdao in China, Humanities and Social Sciences Key Project No. XNR201601.

Xylogiannopoulos K. et. all,(2017), “Text Mining in Unclean, Noisy or Scrambled Datasets for Digital Forensics Analytics”, 2017 European Intelligence and Security Informatics Conference, 978-1-5386-2385-5/17 $31.00 © 2017 IEEE DOI 10.1109/EISIC.2017.19.

Andriansyah M. et. all, (2014), "Developing Indonesian Corpus of Pornography using Simple NLP-Text Mining (NTM) Approach to Support Government Anti-Pornography Program"

Harshni S. et. all, (2017), "Data Mining and Text Mining – A Survey", 2017 International Conference on Computation of Power,Energy, Information and Communication (ICCPEIC), 978-1-5090-4324-8/17/$31.00 ©2017 IEEE.

Alshuwaier F. et. all, (2017)"A Comparative Study of the Current Technologies and Approaches of Relation Extraction in Biomedical Literature using Text Mining", 978-1-5090-4324-8/17/$31.00 ©2017 IEEE.

Muhamad S. et. all, (2017), "Cyberbullying Classification using Text Mining" , 2017 1st International Conference on Informatics and Computational Sciences (ICICoS), 978-1-5386-0903-3/17/$31.00 © 2017 IEEE.

Ghosh S. et. all, (2012), "A tutorial review on Text Mining Algorithms", *International Journal of Advanced Research in Computer and Communication Engineering*, Vol. 1, Issue 4, June 2012, ISSN : 2278 – 1021.